**In: Rudolf Muhr, Eugênia Duarte, Amália Mendes, Carla Amóros Negre, Juan Testa (eds.) (2015): Pluricentric Languages and non-dominant Varieties worldwide: Pluricentric Languages across continents - Features and usage. Wien et. al., Peter Lang Verlag. p.**

*Amália MENDES, Antónia ESTRELA, Fernanda BACELAR DO NASCIMENTO, Luísa PEREIRA, Sandra ANTUNES*

(Centro de Linguística da Universidade de Lisboa, Portugal)
amalia.mendes@clul.ul.pt, antoniaestrela@gmail.com, fbacelar.nascimento@gmail.com, luisa.alice.sp@gmail.com, sandra.antunes@clul.ul.pt

# *New words, old suffixes: Nominal derivation in the African varieties of Portuguese compared to European Portuguese*

***Abstract*** *(Maximum 150 words/900*

The aim of this paper is to offer an account of the nominal suffixation patterns found in the African varieties of Portuguese, under a contrastive study with the European Portuguese norm. Our study is based on data from the Corpus Africa, contrasted with the lexicon extracted from the Reference Corpus of Contemporary Portuguese, as well as lexicographic works of reference. Each subcorpus shows innovation in nominal suffixation (compared to the European norm) and these innovations occur in one, at most two or three AVP. These new words can be divided in two types: (i) words that have no equivalent with the same lexical basis in EP and (ii) words that have a lexical equivalent in EP, but use a different suffix in its formation. We will clarify the role that suffixation plays in the formation of new nouns in AVP, regarding the morphological and semantic properties of the basis selected by suffixes with equivalent meaning.

## 1. *Introduction*

With this paper, we intent to offer an account of the nominal suffixation patterns found in the African varieties of Portuguese, under a contrastive study with the European Portuguese norm. We will base our study on data from the Corpus Africa – which includes subcorpora of the five African countries where Portuguese is the official language

and is actually used by the population: Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe (Bacelar do Nascimento, 2006). These data are contrasted with lexicographic works of reference and with the Reference Corpus of Contemporary Portuguese, restricted to the country "Portugal" (Généreux et al, 2012), in order to assess the diversity encountered when compared to Portugal. Quite different situations are encountered in the Portuguese speaking countries: while it is a crucial factor of national unity in Portugal and also in Brazil, it is frequently acquired as a second language in the 5 African countries in our study. The wide dissemination of Portuguese in Brazil, its dimension in terms of number of speakers and the economic and political importance of Brazil in the new international geopolitical order accounts from the ongoing shift from an exonormative perspective to an endonormative one (Stewart 1968). The situation is still quite different in the African countries: in Cape Verde, Guinea-Bissau and S. Tome and Principe, Creole languages are widely used and Portuguese is spoken by a minority; in Angola and Mozambique there are no Creole languages and the use of Portuguese has in fact been increasing in the recent years, as a factor of national unity in a multilinguistic environment. In these 5 countries, Portuguese is in a very strong situation of language contact and it happens that this language is often spoken as L2 (cf. Gonçalves 1990: 272) or is acquired as a first language from input of speakers of Portuguese L2. This is frequently considered to be the motivation for the diverging phenomena encountered in these varieties. To test this hypothesis, we will compare our data with cases of noun formation that diverge from the European norm in COPLE2, a learner corpus of Portuguese (Antunes et al. 2015). We are certainly aware that these data are quite different, in the sense that COPLE2 is a corpus of learners of Portuguese as a foreign language, where no immersion in a Portuguese-speaking environment has taken place, contrary to most of the informants of the corpus of African varieties of Portuguese. With all this in mind, we still believe that the comparison may prove fruitful to establish to what degree do nominal sufixation processes differ from the European norm, on the one hand, and to what degree are they comparable to noun formation by foreign learners of Portuguese.

The study will be organized as follows. We will initially present the description of the CA, with respect to its constitution and the

organization of the five lexicons (section 2). Then, in section 3, we will describe the methodology regarding the selection of the lemmas to be studied. In section 4, we present the analysis of the lemmas; in 5, we analyze data from the COPLE corpus. We conclude in section 6, with some relevant points related to the suffixation in AVP.

## 2. *The Corpus Africa*

The Corpus Africa (CA) has around 640.000 words for each variety, including spoken and written texts, and the five corpora are comparable in size, chronology and broad types and genres. (c. 25,000 spoken words [4 %] and c. 615,000 written words). For a detailed description of the corpus, its dimension per variaty and design, cf. Bacelar do Nascimento *et al.* (2006). The CA has been lemmatised and automatically annotated with Parts-of-Speech information and inflection using Eric Brill's tagger (Brill 1993), trained over the PAROLE corpus, a written and spoken Portuguese corpus of 250,000 words, morphosyntactically annotated and manually revised (Mendes et al., 2004).

Five lexicons have been extracted from the corpora, comprising lexical items from the main categories: Common Noun, Adjective and Verb. These lexicons have been compared and treated statistically in the form of contrastive lists that provide information on the set of lemmas that constitute the common core vocabulary of the five subcorpora, but also on the set of those that only occurred in one or in some of the subcorpora. The vocabulary common to all the subcorpora is considered to be the "core lexicon" and corresponds to 26% of the total set of lemmas in the corpus, while the lemmas that occurred in just one of the corpora, i.e., the peripheral lexicon (Nelson, 2006) corresponds to 37% of the lemmas. However, the 26% of the lemmas of the core vocabulary corresponds to 91.75% of occurrences in the corpus, while the lemmas of the peripheral lexicon present low frequencies or are hapax legomena, and are, in fact, more representative cases of lexical change, or africanization and provide important clues into the lexical change

undergone by different varieties of Portuguese (cf. Baclear do Nascimento et al., 2008 for more details).

### 3. *Contrastive analysis of nominals derived by suffixation*

Our study focuses on nominal forms that are the results of suffixation processes in the CA. The first task was to identify all nouns formed by suffixation in the corpus and to see if they were common to the European Portuguese variety or specific of the African varieties, and if they follow regular morphological processes of suffixation or not. We take the European Portuguese norm as a reference for our study, while being attentive to the actual word forms encountered in European Portuguese usage. We compared the lexicon extracted from the CA corpus with lexicographic reference works: an online dictionary of Portuguese: *Dicionário Priberam da Língua Portuguesa*: and a printed dictionary of reference: *Dicionário da Língua Portuguesa* Porto Editora, 2003. All the words encountered in these lexicographic references (with no special information on their variety) were considered as used in the European variety of Portuguese and were consequently excluded from our study. However, we also excluded words that did occur in the dictionaries when they were labelled as *africanism*, i.e., words that were imported from African languages, such as *anhara, ambundo, badjuda* 'girl', since they are not cases of derivation in Portuguese and don't fall into the topic of our study. We then compared the remaining list with the oficial orthographic vocabulary of Portuguese available online (*Vocabulário Ortográfico do Português (Portal da Língua Portuguesa)*), which includes information on the Portuguese varieties and isolated those that were not encountered, or that were attributed to an African variety. This produced a list of derived nominals that were found in the Corpus Africa and not in lexicographic references. This list was furthermore compared to the *Reference Corpus of Contemporary Portuguese*, by selecting a restricting query over the subcorpus "Portugal". The corpus provides important information regarding the usage of some words that may not be found in the dictionaries. It is worth mentioning that our intuition about what would be a non-European Portuguese word form turned out to be frequently misleading since many of such cases turned out to be categorized as

archaic word forms in the dictionaries of Portuguese, and this is the kind of information that the sole use of general synchronic corpora wouldn't provide.

There is a total of 25,523 lemmas in the CA and, after inspection of lexicographic references, a total of 174 lemmas were retained as candidates for our study: those are the lemmas that are considered specific to the CA, after comparison with the 3 lexicographic references. After inspection of the CRPC, 31 lemmas that were encountered in the corpus were furthermore excluded from our listing (cf. section 4). The final result is a set of 143 nominal lemmas (and 241 word forms) that were selected as being specific of the corpus of the African varieties. In fact, 107 lemmas, out of the 143, occur a single time in the CA, so that hapax legomena account for most of the nominal forms under study. These 143 nominal forms are distributed as follows in the CA: Angola (73), Cape Verde (62), Mozambique (41), Guinea-Bissau (37) and Sao Tome (28). Most are part of the peripheral vocabulary (specific to 1 variety). In sum, there are relatively few word forms that are specific to the African varieties, when compared to the total number of lemmas, and this challenges the idea that the African varieties differ greatly from the European one in what concerns suffixation processes at least. However, each variety shows innovation in nominal suffixation (compared to the European norm), as illustrated in section 5.

### *3.1 The CRPC as an exclusion corpus*

Although the lemmas that were found in the subpart "Portugal" of the CRPC corpus are not part of our final selection, we believe that a discussion of the exclusion role of the corpus is worthwhile. A set of 31 lemmas were excluded from our selection due to their occurrence in the CRPC subpart "Portugal" and the corresponding 107 word forms are distributed has follows in the CA: Angola (15), Cape Verde (79), Mozambique (5), Guinea-Bissau (8) and Sao Tome (5).
The high frequency of found in Cape Verde corresponds, in fact, to two very frequent words in the Cape Verdean subcorpus: Mindelense 'from Mindelo' occurs 36 times (Mindelo is a city in the island of São Vicente in Cape Verde) and orçamentação 'budgeting', which occurs 19 times.
Some of the lemmas that occur in the CA and in the CRPC-Portugal

corpus, but not in the lexicographic references are in many cases specialized word forms, such as *caseirismo* 'act of favouring a club' (1a), frequently used in football news when refering the referee's performance, or *semanticidade* 'semanticity' in linguistics (1b). Other cases are non specialized lemmas that might be integrated in dictionaries in the future, as the lemma *rendista 'person who pays a rent'* (1 occurrence in CA and 21 in CRPC), *judicialização* 'turning an activity susceptible of judicial action' (1 occurrence in CA and 75 in CRPC) and *orçamentação* 'budgeting' (22 occurrences in CA and 146 in CRPC). Of course, the frequency of occurrence in both corpora is not comparable and is only provided to give an indication of the tendency of use of the lemmas in the different varieties of Portuguese.

In some cases, the word form that occurs in the CA may be in fact more frequent, both in the CA and in the CRPC, than the word form that is listed in the dictionaries. For instance, *abeberação* 'the act of giving water to the animals' appears in the dictionary but not in the corpora, while its synonym *abeberamento* is found 45 times in the CRPC and ?? times in the CA. So, what might seem to be specific to the CA, when taking only into consideration the dictionaries, turns out to be common to the corpora of the African and European varieties of Portuguese.

Furthermore, two lemmas for the same concept may have been formed with different suffixes and be both attested in the CA and in the CRPC, although only one of these lemmas will be dictionarized and have higher number of occurrences in both corpora. Some examples are provided in Table 1, where the least frequent lemma is marked in bold. Notice however that in the CA *desalfandegação* is amost as frequent as its synonym.

| | CA | CRPC-Portugal |
|---|---|---|
| acomodação 'complacency' | 23 | 403 |
| **acomodamento** 'complacency' | 1 | 43 |
| **agrupação** 'grouping' | 1 | 3 |
| agrupamento 'grouping' | 42 | 4284 |
| **desalfandegação** | 4 | 6 |
| desalfandegamento | 5 | 26 |

| **espontaneísmo** 'spontaneity' | 1 | 17 |
|---|---|---|
| espontaneidade 'spontaneity' | 8 | 542 |

Table 1: Pairs of derived nominals in the CA and in the CRPC-Portugal

While "agrupação" occurs 1 time in CA and 3 in CRPC-EP, "agrupamento" occurs 42 times in CA and 4284 in CRPC. In fact, we would only expect the lemma "agrupamento" to occur in CRPC-EP, as it is the only one that is present in the dictionary, but the alternate word has also 3 occurrences in CRPC-EP. This confirms the fundamental role of corpus when describing language varieties.

The corpus provides complementary data to the lexicographic references in what concerns the usage of specialized nominals, as well as concurrent word forms.

### *3.2 Specific lexicon of the Corpus Africa*

We will now focus specifically on the lemmas that are not listed in the lexicographic references, nor found in the subset Portugal of the CRPC. These will, in principle, be the best candidates to provide input to the study of the suffixation processes in place in the African varieties of Portuguese. We will divide our discussion in cases of words formed by regular processes of suffixation (section 3.2.1) and words with irregular suffixation processes (section 3.2.2).

#### 3.2.1 Words formed by regular suffixation processes

We are taking into consideration, in this subsection, 87 of the 143 selected lemmas. Table 2 provides information on the most productive suffixes that contribute to word formation in these lemmas and some examples of lemmas in the CA. In what concerns the most productive suffixes, they are *–ção, -mento, -idade, -eiro, -dor, -ice, -agem* e *–ista*.

| Suffix | Lemmas |
|---|---|
| -ção | *alertação* 'alert', *defendição* 'defence', *destacação* 'secondment', *roncação* 'snoring', *discursivização* 'speech', *angolanização* 'to become Angolan' |
| -mento | *anestesiamento* 'anaesthesia', *evisceramento* 'evisceration', *vigiamento* 'keep watch', *concessionamento* 'granting' |

| | |
|---|---|
| -idade | *facialidade* 'related to the face of something', *minusculidade* 'property of being minuscule', *vincularidade* 'binding', *vinculatividade* 'binding/mandatory', *zambezianidade* 'typical of Zambeze river', *metricidade* 'metre' |
| -eiro | *brilheiro* 'shining', *tartarugueiro* 'someone who takes care of turtles' |
| -dor | *cronicador* 'chronicler', *mestiçador* 'crossbreeder' |
| -ice | *mulatice* 'being mullato', *parlamentice* 'things that happen in the Parliament' |
| -agem | *quilatagem* 'caratage', *mulatagem* 'crossbreeding', *xaropagem* 'syrop', *farinhagem* 'flour' |
| -ista | *Vilista* 'someone from Vila Clotile club', *saxtenorista* 'tenor saxophone', *vandalista* 'vandal', *lunarista* 'someone who explains things based on the moon' |

Table 1: Productive suffixes in the lexicon that is specific to the CA

Lemmas that are specific to the CA have frequently a lexical base of African origin that has been adapted to Portuguese and adjoined a suffix, following regular processes of word formation. Some examples are provided in (1). The lemma *balantização* 'turning into Balanta [the most important ethnic group in Guinea-Bissau]' is formed by nominalization of the verb *balantizar*, which is a verbalization of the African base *balanta*: [balant+iz+a]+ção). The lemma *chambocada* 'a beating with a stick' comes from the ningué word *chamboco* (Mozambique): [chamboc(a)+ada], and *muatismo* 'the property of being a chief' from the niungue word *mwata* 'chief' [muat(a)+ismo].

(1) a. A **balantização**, ou o fomentar inconsciente do tribalismo na Guiné-Bissau"(GB)[1] 'the balantization, or the unconscious incentive to tribalismo in Guinea-Bissau'

b. Sou seu pai. E ditas as três palavrinhas desfechou uma matraca sobre o outro. Uma, duas, quatro **chambocadas**. As suficientes, mortais." (MO) 'I'm your father. And after these 3 little words he stroke him with a matraca. One, two, four

[1] For each example of the corpus, the variety of Portuguese is indicated with a two letter code: AN (Angola), CV (Cape Verde), GB (Guinea-Bissau), MO (Mozambique) and ST (Sao Tome and Principe).

chambocadas. The sufficient ones, mortals.'

c. vamos acabar com os privilégios dos responsáveis, com o **muatismo**. (AN) 'let's end the priviledges of the ones in charge, the muatismo'

The lemmas might not be formed over an African lexical base and still reflect an entity or situation that denotes specific aspects of the African reality. It is the case of lemmas that refer to a specific country in Africa, e.g. *angolanização* 'to become Angolan', *antiangolanismo* 'anti-Angolan', *burundês* 'from Burundi', *zambezianidade* 'typical from Zambeze'. And of lemmas where the base is an acronym for a political party which forms the lexical base for the new suffixation, as illustrated in (2) with UCID [UCID+ista] and FLQ [FL(e)Qu+ista] 'member of the UCID / FLQ party'. In the case of the acronym FLQ, the new word form involves the insertion of a vowel to produce a syllabic structure in Portuguese.

(2) a. Não estou na **UCID** de ânimo leve (...) Mais recentemente, certas doninhas fedorentas, com a capa de **Ucidistas**, deitaram a unha de fera, mas eu me mantive calado (...). (CV) 'I'm not in the UCID [party] light-heartedly (...) More recently, certain stinking weasels, pretending to be Ucidists, attacked, but I kept quiet.'

b. Tenta também reagrupar a **FLQ** à sua volta em Paris. Vários **flequistas** no exílio encontram-se nessa época em França. (CV) '[He] tries to regroup FLQ around him in Paris. Several members of FLQ were in France at the time.'

Some of the word forms are specialized lexicon that isn't listed in the dictionaries nor do they appear in the CRPC. This is a similar situation to the discussion of the lemmas *caseirismo* and *semanticidade* in section 3.1. and differ only because in this case the lemma doesn't occur at all in the Portugal subset of the CRPC. It is the case of the word *aburação* that is found in a legal text in the CA (cf. (3)). A query over the internet returns a few hits with the word form, in European Portuguese legal texts.

(3) Não é permitido ao Gerente obrigar a sociedade em actos ou contratos interesse à ele alheios designadamente em fianças,

**aburações**, letras de favor e outros actos semelhantes." (ST)

We discussed in 3.1 cases of two concurrent synonym forms with two different suffixes, illustrated in Table 1. Both lemmas occurred in the CRPC and in the CA, while one was clearly less frequent and not listed in the dictionaries. In the cases illustrated in (4), the situation is slightly different since one of the forms is absent from the CRPC corpus and only occurs in the Africa corpus. These are cases where one of the word forms isn't attested in European Portuguese but do show up in one or more of the corpora of African varieties. For instance, *destacação* doesn't occur in CRPC-EP, but *destacamento* has 797 occurrences.

| | CA | CRPC-Portugal |
|---|---|---|
| **evisceração** 'evisceration' | 7 | 0 |
| evisceramento 'evisceration' | 1 | 1 |
| plantação 'planting' | 60 | ?? |
| **plantagem** 'planting' | 1 | 0 |
| **destacação** 'secondment' | 1 | 0 |
| destacamento 'secondment' | 23 | 797 |

Table 3: Pairs of derived nominals in the CA and in the CRPC-Portugal

These data show that there is more hesitation and variation in the CA regarding which form (and which suffix) should be used than in the European Portuguese corpus, where suffixed nouns seem to be more established and follow more closely the lexicographic references.
There seems to be a tendency in the CA for word formation by suffixation in cases where a concurrent form created by conversion is listed in the dictionary and is frequent or highly frequent both in CRPC corpus and in CA. We provide some exemples of such cases in Table 4. Notice however that such word forms have extremelly low frequencies in the CA corpus. Nevertheless they do not occur at all in the CRPC-Portugal corpus and this again points to some hesitation regarding how to produce such nominals in the African varieties of Portuguese.

| | CA | CRPC-Portugal |
|---|---|---|
| **alertação** 'alert' | 1 | 0 |
| alerta 'alert' | 72 | |
| **roncação** 'snoring' | 1 | 0 |
| ronco 'snoring | 7 | |
| **queimança** 'burning' | 1 | 0 |
| queima 'burning' | 23 | |
| **defendição** 'defense' | 1 | 0 |
| defesa 'defense' | 589 | |
| **anestesiamento** 'anesthesia' | 1 | 0 |
| anestesia 'anesthesia' | 6 | |

Table 4: Pairs of derived nominals in the CA and in the CRPC-Portugal

### 3.2.2 Words formed by irregular suffixation processes

Our discussion in 3.2.1 was centred on lemmas that are specific to the Corpus of African varieties of Portuguese and that follow regular processes of lexical formation. However, we also encounter in our data cases of word formation through suffixation processes that don't follow morphological rules of derivation or the selection properties of the suffixes.

This is the case when the suffix is adjoined to a lexical base pertaining to a part of speech category not selected by the suffix. There is a total of 14 (out of 143) such cases in the CA. For instance, the suffix –idade, which is a deadjectival nominalizer suffix that combine with adjectival bases to form a nominal form, is adjoined in (4a) to the nominal unit *planície* to form the word planicidade [planíci(e)+idade] 'property of being plain'. A similar case is presented in (4b), where the suffix combines with the nominal unit *arbor* (a latinism) to form the noun arboridade [arbor+idade] 'property of being a tree'.

(4) a. A Chã das Caldeiras constitui uma imensa caldeira cujo diâmetro maior atinge nove quilómetros, onde é suave a topografia, sendo a **planicidade** interrompida por alguns cones

adventícios e cordões de lavas." (CV) 'The Chã das Caldeiras place is a huge cratera whose larger diameter is nine quilometers, where topography is smooth, and the "plainess" is interrupted by some XX cones and XX'

b. Platão chamou a nossa atenção para um aspecto genérico da linguagem, de que um determinado substantivo ou adjetivo, por exemplo, 'árvore' ou 'agudo', pode ser verdadeiramente aplicado no mesmo sentido a um grande número de coisas distintas e diferentes: a sua opinião é de que isso só será possível se existir alguma entidade designada pelo termo geral em questão o **arboridade**, agudeza - da qual compartilha cada um dos indivíduos. (AN) 'Plato called our attention to a generic aspecto of language, that a noun or adjective, for instance 'tree' or 'sharp' may be applied with the same meaning to a large number of distinctive and diferente things: his opinion is that it could only be possible if there is some entity called by the general term in question, treeness, sharpness, shared by each individual'

Another example is provided by the suffix –ice, which, according to the norm, combines with adjectives to form nouns. The suffix occurs in the CA combined with a verbal base, as illustrated in (5a) with the lemma *atrapalhice* [atrapalh(ar)+ice] 'fumbling'. The synonym formed with the suffix –ção (*atrapalhação*) is well formed and used in the CA (freq. 5) and the CRPC-Portugal (freq ). Another exemple is the lemma *chaleirice* 'flattery', formed over the verbal base *chaleirar*, marked in the dictionary as being used in the Brazilian variety [chaleir(ar)+ ice]

(5) a. o chefe da secção, na **atrapalhice** das pressas, tinha chegado primeiro" (DO) 'the chief of the section, in the confusion of the hurry, arrived first'

b. Disse mais que só agora, depois de me ver, compreendeu que o seu destino se fixou... - E tu, é claro, acreditaste nas suas **chaleirices**... (CV) 'And [he/she] said that only now, after seeing me, [he/she] understood that her destiny was settled... - Andy ou, of course, believed in his/her flattery'

Cases illustrated in (4) and (5) do not follow the norm in terms of the selection properties of the Portuguese suffixes. These specific lemmas were not encountered in the CRPC-Portugal, but it is important to consider wether other lemmas might have been formed in similar ways in the European and Brazilian varieties of Portuguese. In fact, the combination of the suffix –ice with verbal lexical bases seems to be productive both in EP and BP in cases such as *bajulice* 'flattery' [bajul(ar)+ice] and *choraminguice* 'whimper' [choraming(ar)+ice]. The fact has been already noted over data from a Brazilian Portuguese corpus in Pezatti (1990:157) for the suffix –idade "The data also show that this morpheme doesn't always combine with adjectives, contrary to what is stated in grammars: there are also cases of affixation to a noun or a numeral, as in the examples: ânsia (N) + idade = ansiedade (...) dúplice (Numeral) + idade"[2]. The author also provides examples of the combination of the suffix –ice with a verbal theme, such as *alcovitice* [alcovitar+ice] and *coscuvilhice* [coscuvilhar+ice] (Pezatti 1990:168).

Some occurrences in the CA corpus are difficult to analyse in terms of their internal morphological structure. For instance, the lemma *transibilidade* might have been formed by syncope of the segment –ta- in *transitabilidade* 'accessibility', and *obsolência* by syncope of –esc- in *obsolescência* 'obsolescence'. Again, this does not affect the suffix itself but the lexical base that it combines with.

### 3.3.3. Discussion of results

We found a restricted list of lemmas exclusive to the CA corpus. Most of these lemmas follow regular patterns of word formation. They are specific to the CA corpus due mostly to the fact that they are formed over a lexical base of African origin, or denote a specific reality of the African countries. Some of the lemmas are concurrent forms of a highly frequent synonym that occurs both in the CA and in the CRPC-Portugal. This indicates a linguistic situation where the speakers have a deep knowledge of the morphological rules of suffixation and produce possible and grammatical nominal forms, although not attested in

[2] Our translation. Original: "Os dados mostraram ainda que nem sempre tal morfema se agrega a adjetivos, como atestam as gramáticas: há também casos de afixação a substantivo e a numeral, conforme os exemplos (...)".

dictionaries and in concurrence with other well established forms (CorreiaXXX). It does also provide evidence of some hesitation in what concerns the choice of suffix, grounded in a less proficient lexical knowledge that might relate to the status of Portuguese as a second language or as a first language acquired in an environment of speakers of Portuguese L2.

Non regular patterns of suffixation are mostly related to a wrong categorial choice of the lexical base, although exceptions found with some suffixes are also encountered in EP. It is important to take into consideration corpus data in order to put under perspective the European norm that is established in grammars and other reference works and that might lead us to consider such word formation processes as specific of the African varieties of Portuguese.

## 4. *Comparison with acquisition data of Portuguese FL/L2*

### *4.1 COPLE2 – Corpus of Portuguese FL/L2*

COPLE2[3] is a new corpus of Portuguese as a foreign language (FL) or second language (L2), which encompasses written and spoken data produced by foreign learners of Portuguese at the University of Lisbon between 2010 and 2012. This corpus aims at providing empirical data for the teaching and learning of L2 Portuguese by: (i) identifying general errors in the learning of Portuguese FL/L2 (Granger, 1996); (ii) developing textbooks and other teaching material targeting students with specific mother tongues (L1).

Our analysis is based on data from the written register of COPLE2, which is composed by:

(i) 966 free handwritten essays from different genres (dialogue, formal and personal letters, informative, message/e-mail, argumentative, recount, book review), that were collected in evaluation tests or accreditation exams, in a total of 156.691 tokens;

(ii) 424 students aged between 18-40 years;

[3] http://www.clul.ul.pt/en/research-teams/547

(iii) 14 different L1s (Chinese, English, Spanish, German, Russian, French, Japanese, Italian, Dutch, Tetum, Arabic, Polish, Korean and Romanian);
(iv) all levels of proficiency, according to the levels of the CEFR[4] (Beginner (A1), Elementary (A2), Intermediate (B1), Advanced (B2), and Proficient (C1)).

We restrict our analysis to learners of Portuguese with Spanish and English as L1s (cf. Table 1). We choose these different languages (a Romance and a Germanic language) to see if the students' L1 play a significant role when they have to deal with suffixation in Portuguese.

| L1 | Inf. | Texts | Total Words | Words /Text |
|---|---|---|---|---|
| English | 65 | 142 | 21.610 | 152 |
| Spanish | 52 | 139 | 21.200 | 153 |
| **TOTAL** | **117** | **281** | **42.810** | **305** |

Table 1: Subpart of the COPLE2 written corpus for the analysis

*4.2. English as L1*

Analysing the production of English students, we found different types of problems regarding the choice of the suffix:

(i) Suffixation over a lexical form with equivalent meaning

a) *patriotista* (patriot(a)+ista) [*patriota* 'patriot']
*Por exemplo, se um brasileiro escrevesse bem português, de qual pais seria* ***patriotista****?* (en053CVCDGF.txt)
b) *poligamista* (poligam(o)+ista) [*polígamo* 'polygamous']
*O vice-presidente Zacob Zuma foi* ***poligamista*** *com mais de 20 filhos* (en028CAAGF_2.txt)

(ii) Part of Speech category selected by the suffix

*actura* (act+ura) [*actuação* 'acting']

[4] http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

*Quando o curso de português acabar eu quero estudo teatro porque eu gosto muito da* ***actura***. (en031CVITF_2.txt)

(iii) Choice of the wrong suffix

*improfessionalidade* (im+profissional+idade) [*pouco profissional* 'unprofessional'

*Nunca tenho visto tão* ***improfessionalidade*** *como isto.* (en050CVCDI.txt)

(iv) Choice of the wrong lexical base

a) *precisável* 'needful' (precis+a+vel) [*prestável* 'helpful'] (there is a semantic relation between *precisar* 'to need' and *prestar* 'to help')

*Os Portugueses são muito simpática e sempre* ***precisável***. (en070CVCDF1_2.txt)

b) *patronismo* [*patriotismo* 'patriotism] (the student used the base *patrão* 'boss' instead of *pátria* 'homeland')

*Na minha opinião, Escrever bem não é fundamentalmente uma forma de* ***patronismo***. (en054CVCDF.txt)

(v) L1 influence

a) *Europeanos* 'Europeans' [*europeus*]
*A maioria dos países* ***Europeanos*** (en036CVEDI.txt)

b) *Govermentos* 'governments' [*governos*]
*o meu pais tem de gastar pouco e poupar muito, não só as pessoas mas o* ***govermentos*** *e cada estado também* (en025CACDD.txt)

(vi) Typos/Errors

a) *moderniçar* [*modernizar* 'to modernize']
*precisamos as novas tecnologias para* ***moderniçar*** *todo o mundo* (en036CVEDF_2.txt)

b) *indespensível* [*indispensável*] 'indispensable'
*Era uma expirência* ***indespensível***. (en070CVCDF2.txt)

*4.3. Spanish as L1*

(i) Lemma formed by suffixation – equivalent lemma formed by conversion

a) *ensinamento* (ensina(r)+mento) [*ensino* 'teaching']
*Tenho feito também imensos cursos relativos ao* ***ensinamento*** (es053CVMTD.txt)

b) *reservação* (reserve(r)+ção) [reserva 'booking']
*Tive que fazer uma nova* ***reservação*** *noutro hotel* (es028CVMTI.txt)

(ii) Lexical base influenced by L1

a) *dificultade* [*dificuldade* 'difficulty'] (from the Spanish *dificultad*)
*viver muito tempo fóra não supõe uma* ***dificultade*** (es032CVATI.txt)

b) *alheamiento* [alheamento 'alienation'] (from the Spanish *alejamiento*)
*uma educação que evite o nosso* ***alheamiento*** *total* (es054CVMTI.txt)

(iii) Typos/Errors

a) *riqueça* [*riqueza* 'wealth']
*Deram dinheiro e* ***riqueça*** *rápida* (es032CVMTD.txt)

b) *miradoio* [*miradouro* 'viewpoint']
*subi ao* ***miradoio*** (es046CVETF_2.txt)

*4.4 Discussion of results*

The analysis of the subsets of COPLE2 points to some interesting preliminary results:

(i) the English subset showed a wider diversity of errors;
(ii) errors due to L1 influence are more frequent in the Spanish subset (a language similar to Portuguese) ;

(iii) errors on the wrong choice of the suffix or the lexical base occurred only in the English subset;
(iv) errors regarding the word formation process only occurred in Spanish (suffixation instead of conversion);
(v) spelling errors over the lexical base or the suffix are frequent in both subsets.

## *5. Conclusion*

A set of unknown word forms have to be further analysed with the input of native speakers of the varieties

The contrastive analysis with the subsets of COPLE2 points to some interesting preliminary results:

- Some processes were found in both corpora, although with limited expression in the COPLE2 corpus, such as:

(i) Wrong choice of categorial lexical base

(ii) Lemma is formed by suffixation while the lemma used in EP is formed by conversion

(iii) Error on the lexical base

- Spelling errors over the lexical base or the suffix are more frequent in COPLE2. Errors due to L1 influence are frequent in the Spanish L1 subset and much less in the English L1 and CA.
- L1 transfers are not comparable: Africanisms were discarded from our analysis, while L1 transfers in COPLE2 are still close to the Portuguese words formed by suffixation

These new words can be divided in two types: (i) words that have no equivalent with the same lexical basis in EP and (ii) words that have a lexical equivalent in EP, but use a different suffix in its formation.

In the production of regular word through the word formation rules there is a wide variety of competing suffixes. These various suffixes have sometimes the same semantic function which can act on the same basis, providing certain irregularity in the process. According to Sandamann (apud Areán-García, 2010), we have here a situation of

conflict between system and use. The system allows more than one form, but the use consecrates and privileges another one.

A change in the suffix and the maintenance of the base not always imply synonymy, and there may be effects of semantic difference. Competition between suffixes can promote regional differences, different levels of formality, status, among others (Areán-García 2010).

*References*

Areán-García, Nilsa (2010): Concorrência entre sufixos: uma visão diacrônica. In I. M. Alves et al. (org.). Os estudos lexicais em diferentes perspectivas. São Paulo: Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. 173-191.

Bacelar do Nascimento, Maria Fernanda, José Bettencourt Gonçalves, Luísa Alice Santos Pereira, Antónia Estrela (2006) "The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon", Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), May 24-26, Genoa, Italy.

Bacelar do Nascimento, Maria Fernanda, Antónia Estrela, Amália Mendes, Luísa Pereira (2008) "On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications", in Zweigenbaum, Pierre et al. (eds.) *Proceedings of the Workshop on Building and Using Comparable Corpora, VI Language Resources and Evaluation Conference - LREC2008,* Marrakesh, Morroco, May 31, 2008, pp. 39-46.

Génereux, Michel, Iris Hendrickx, Amália Mendes (2012) "Introducing the Reference Corpus of Contemporary Portuguese On-Line". In Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC 2012, Istanbul, May 21-27, 2012, pp. 2237-2244.

Gonçalves, Perpétua (1990) A Construção de uma Gramática do Português em Moçambique: Aspectos da Estrutura Argumental dos Verbos. Lisboa. Universidade de Lisboa (PhD).

Granger, S. 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University

Press. Pp. 37-51.Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. (eds.) 2009. *International Corpus of Learner English. Version 2.* Presses Universitaires de Louvain. Belgium.

Mendes, Amália, Raquel Amaro, Maria Fernanda Bacelar do Nascimento (2004) "Morphological Tagging of a Spoken Portuguese Corpus Using Available Resources", in Branco, António, Amália Mendes, Ricardo Ribeiro (eds.) *Language Technology for Portuguese: Shallow processing tools and resources,* Lisboa, Colibri.

Rio-Torto, Graça. (2007) Caminhos de Renovação Lexical: Fronteiras do Possível. In Isquerdo, A. N., Ieda, M. A. (Orgs.), *As Ciências do Léxico, Lexicologia, Lexicografia, Terminologia,* Vol. 3. Campo Grande, MS, Ed. UFMS: São Paulo, Humanitas.

Pezatti, Erotilde Goreti (1990) "A gramática da derivação sufixal: os sufixos formadores de substantivos abstractos", *Alfa* 34, pp. 153-174.